Exascale Begins at the Memory System

Bruce Jacob

Keystone Professor
University of Maryland

OUTLINE

- Problem: Too Much Data Movement & Too Many Memory Accesses
- Reduce the Number (local) → Higher-Capacity Register Files
- Reduce the Costs Per Access → Higher-Bandwidth DRAMs (+ better SerDes)
- Reduce the Number (remote) → Higher-Capacity Main Memory (lower power)
- Reduce the Costs Per Access → Low-Latency, High-Bandwidth Net Topologies

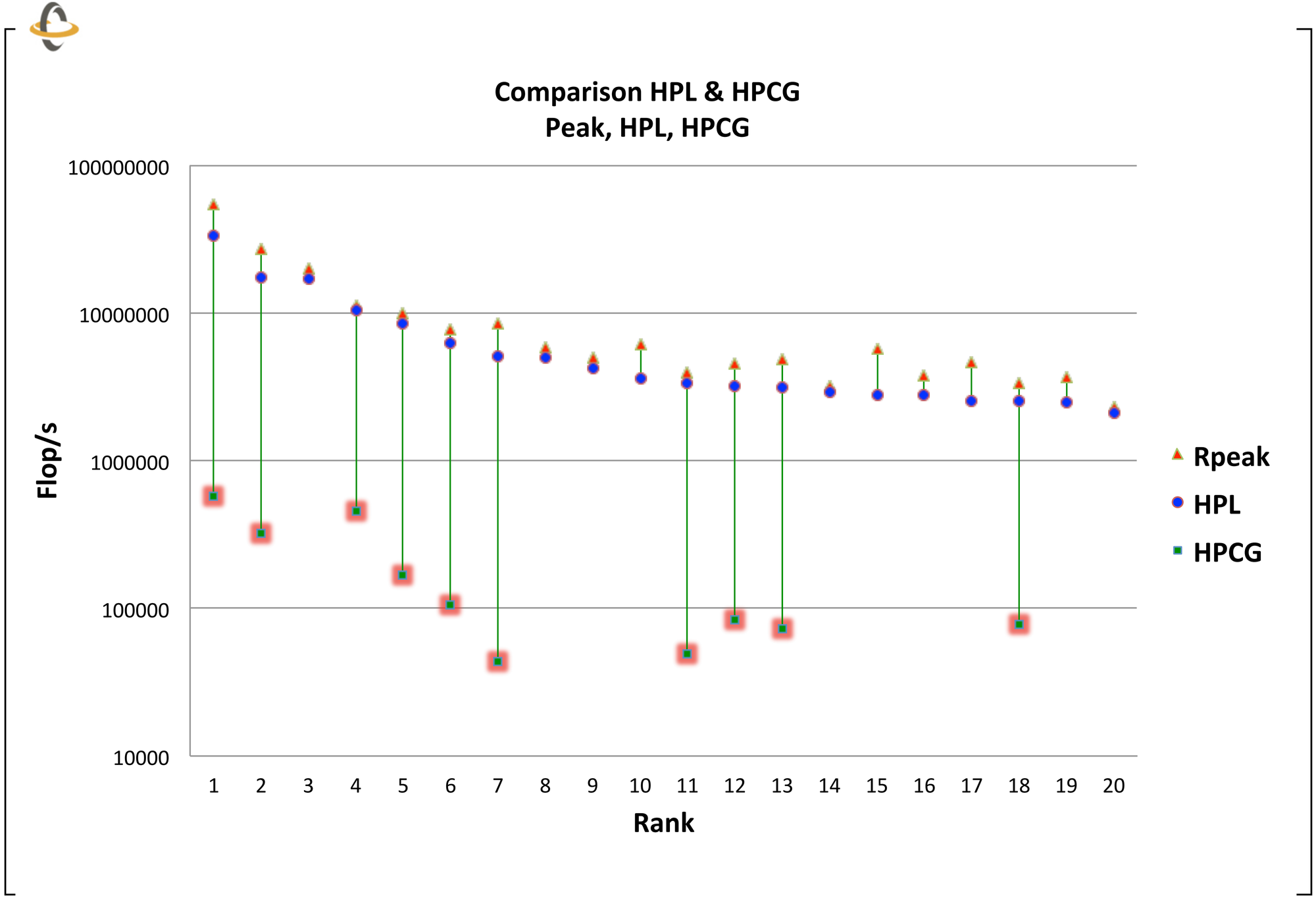
Exascale v. Memory I

Systems sized based on desired data-set, not the number of compute cores needed

Memory can be largest dollar cost *and* power cost

Data movement (esp. all-to-all) truly sucks *(see Jack Dongarra's talk on HPL vs HPCG)*

"90% of supercomputer design is the memory system … and so is the other 10%" — Shekhar Borkar


Comparison HPL & HPCG
Peak, HPL, HPCG
100000000
10000000
1000000
100000
10000
Flop/s
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Rank
Rpeak
HPL
HPCG

Exascale v. Memory I

Systems sized based on desired data-set, not the number of compute cores needed

Memory can be largest dollar cost *and* power cost

Data movement (esp. all-to-all) truly sucks *(see Jack Dongarra's talk on HPL vs HPCG)*

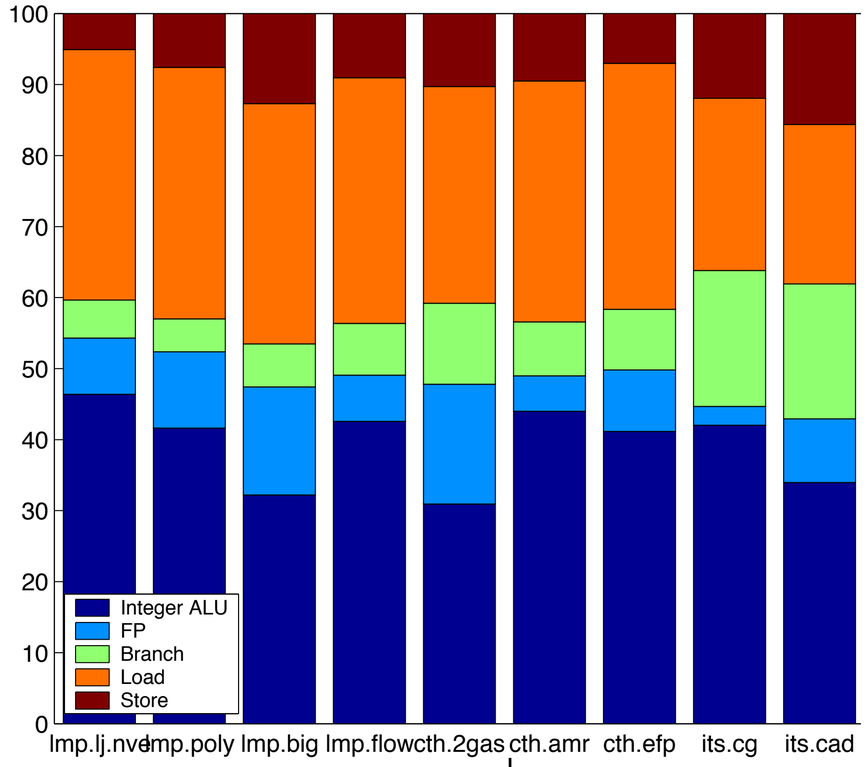
"90% of supercomputer design is the memory system … and so is the other 10%" — Shekhar Borkar

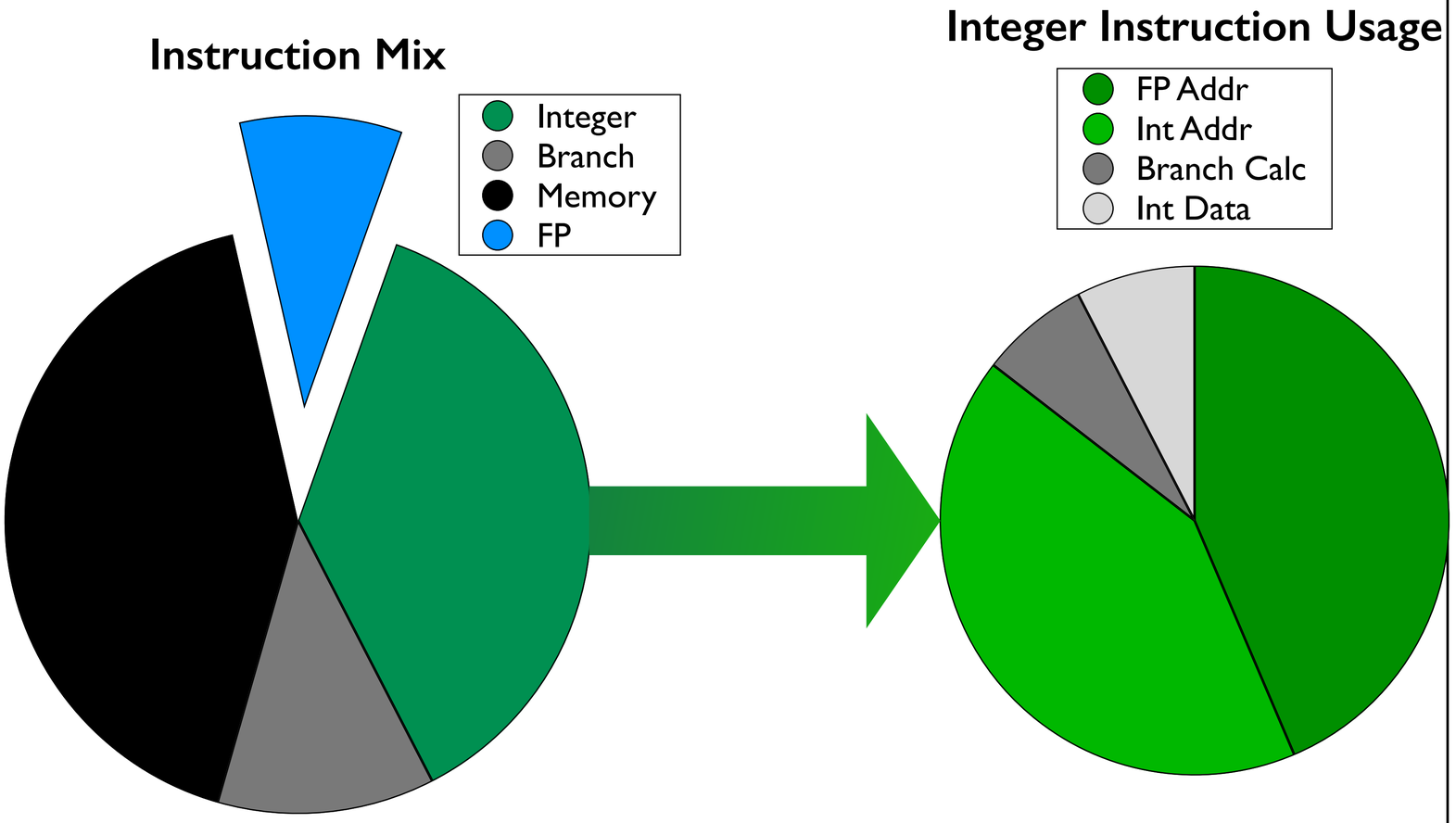
Bruce Jacob

University of
Maryland

Exascale v. Memory II

Sandia Red Storm Instruction Mix





Exascale v. Memory III

Three Quarters of Ops Memory-Related

Likely Sources of Memory Ops:

- **Capacity Limitations** (increase capacity)
- **Unavoidable** (make local access painless)
- **MPI** (reduce frequency of IPC,
  reduce cost of sending data over network)

Exascale v. Memory III

Three Quarters of Ops Memory-Related

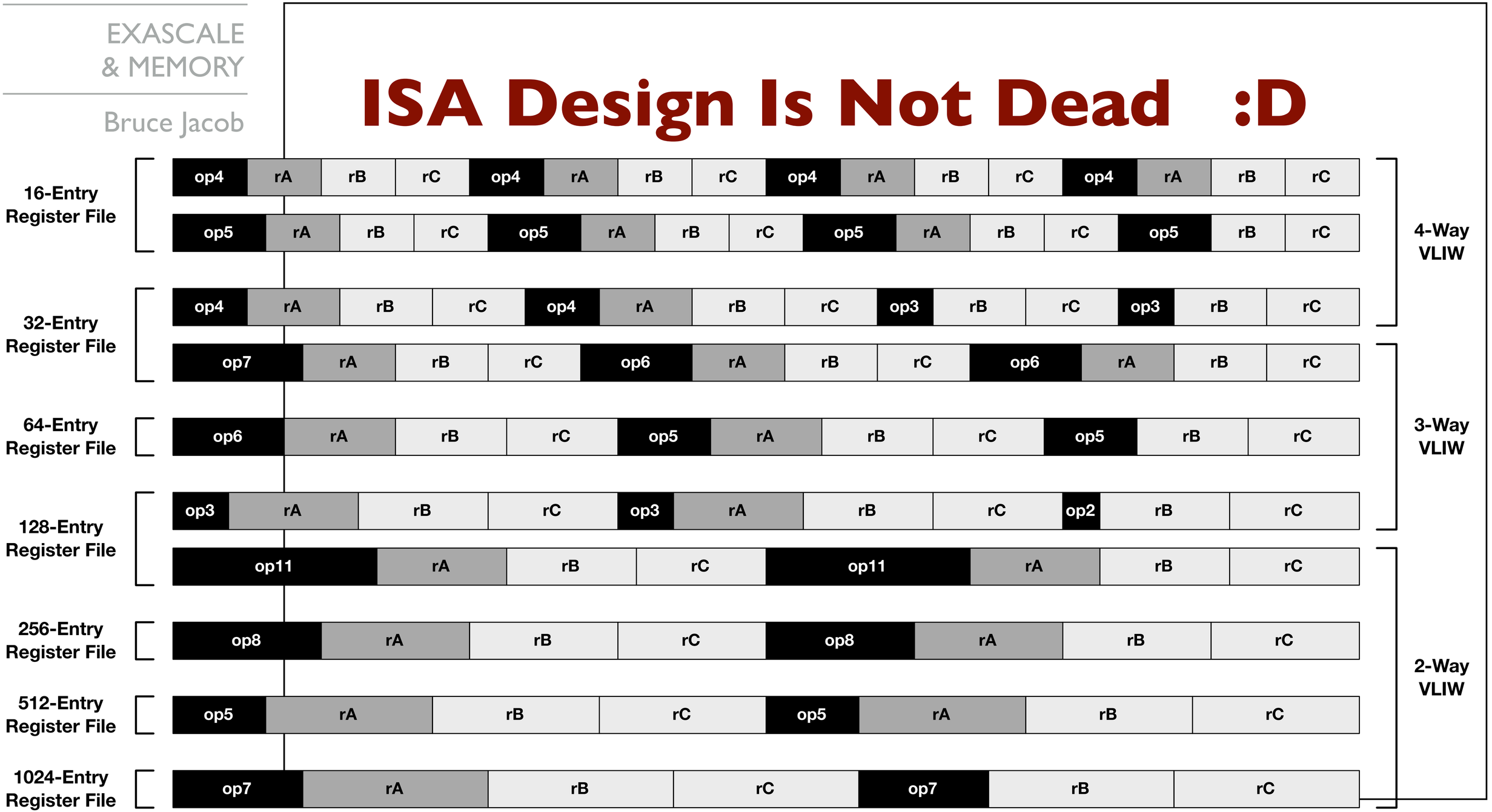
Likely Sources of Memory Ops:

Big RF →

- **Capacity Limitations** (increase capacity)
- **Unavoidable** (make local access painless)
- **MPI** (reduce frequency of IPC,
  reduce cost of sending data over network)

Bruce Jacob

ISA Design Is Not Dead :D

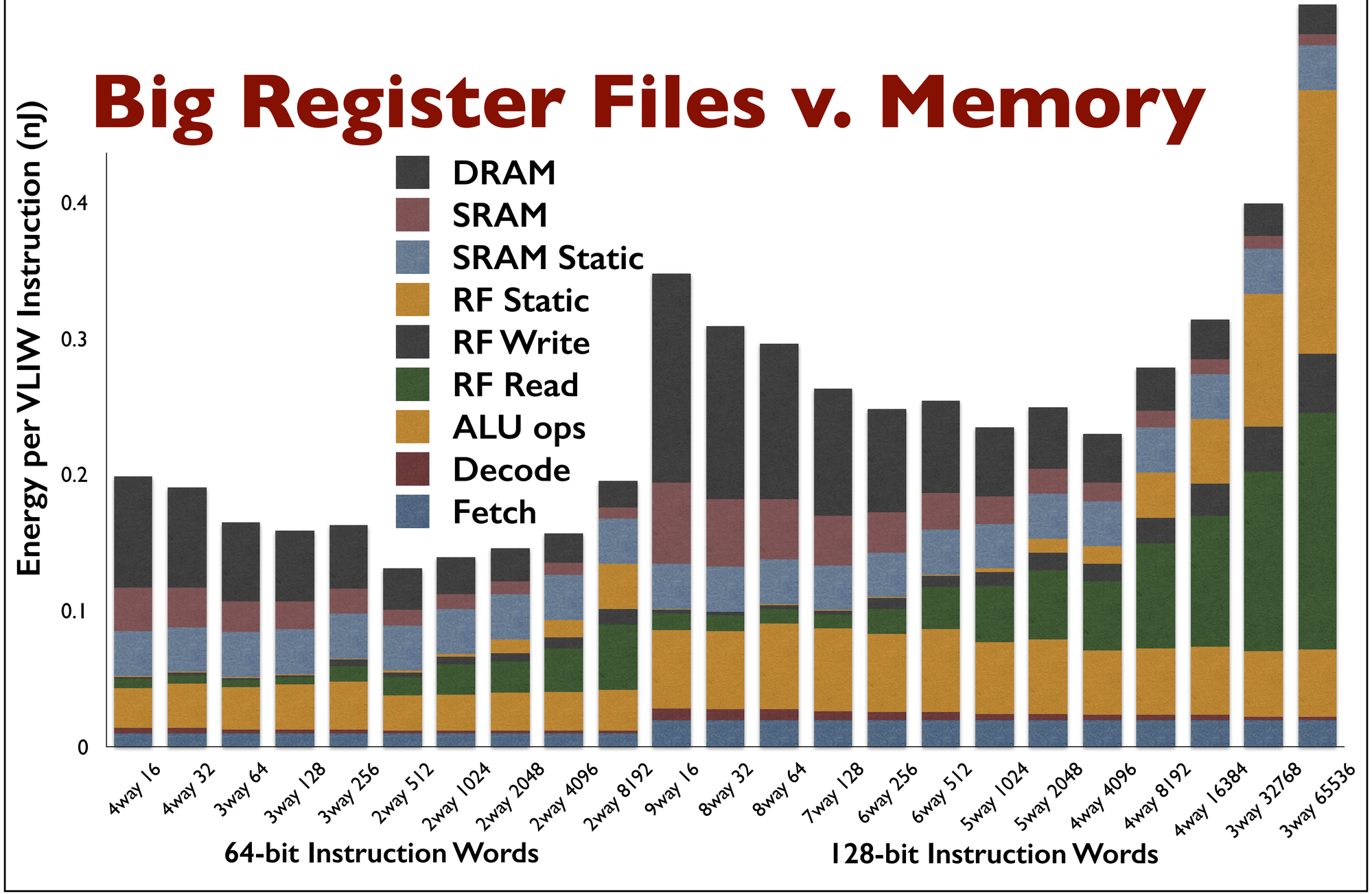


Bruce Jacob

University of
Maryland

Big Register Files v. Memory

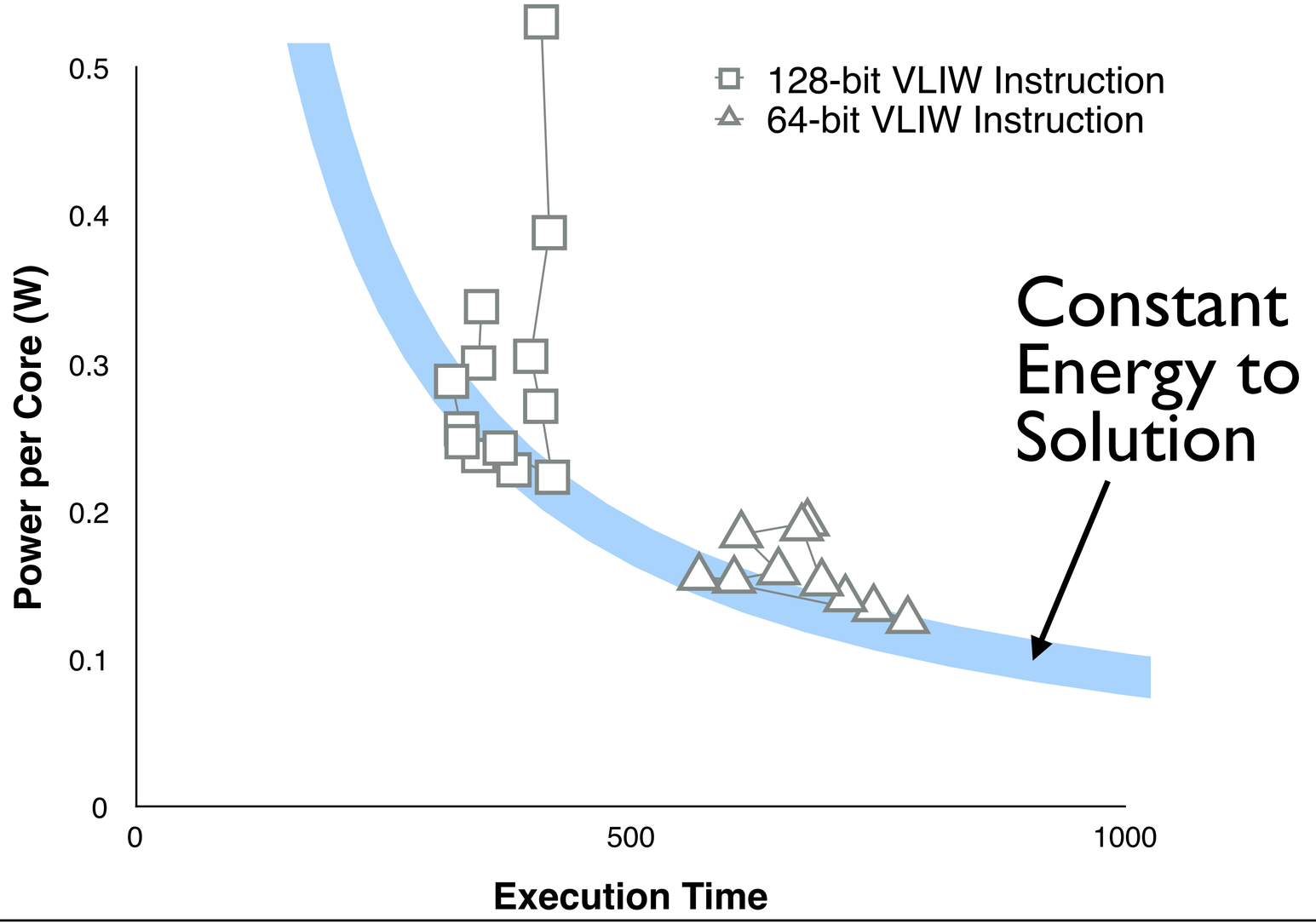
Back-of-the-Envelope Assumptions:
RF @ 90% rule
32 RF → 45% Mem
90/10% SRAM/DRAM
DRAM @ 1pJ/bit, 64B
SRAM/RF via CACTI
Core @ 1GHz (W~nJ)



Bruce Jacob

University of
Maryland

Big Register Files v. Memory



Exascale v. Memory III

Three Quarters of Ops Memory-Related

Likely Sources of Memory Ops:

Big RF →

- **Capacity Limitations** (increase capacity)
- **Unavoidable** (make local access painless)
- **MPI** (reduce frequency of IPC,
  reduce cost of sending data over network)

Exascale v. Memory III

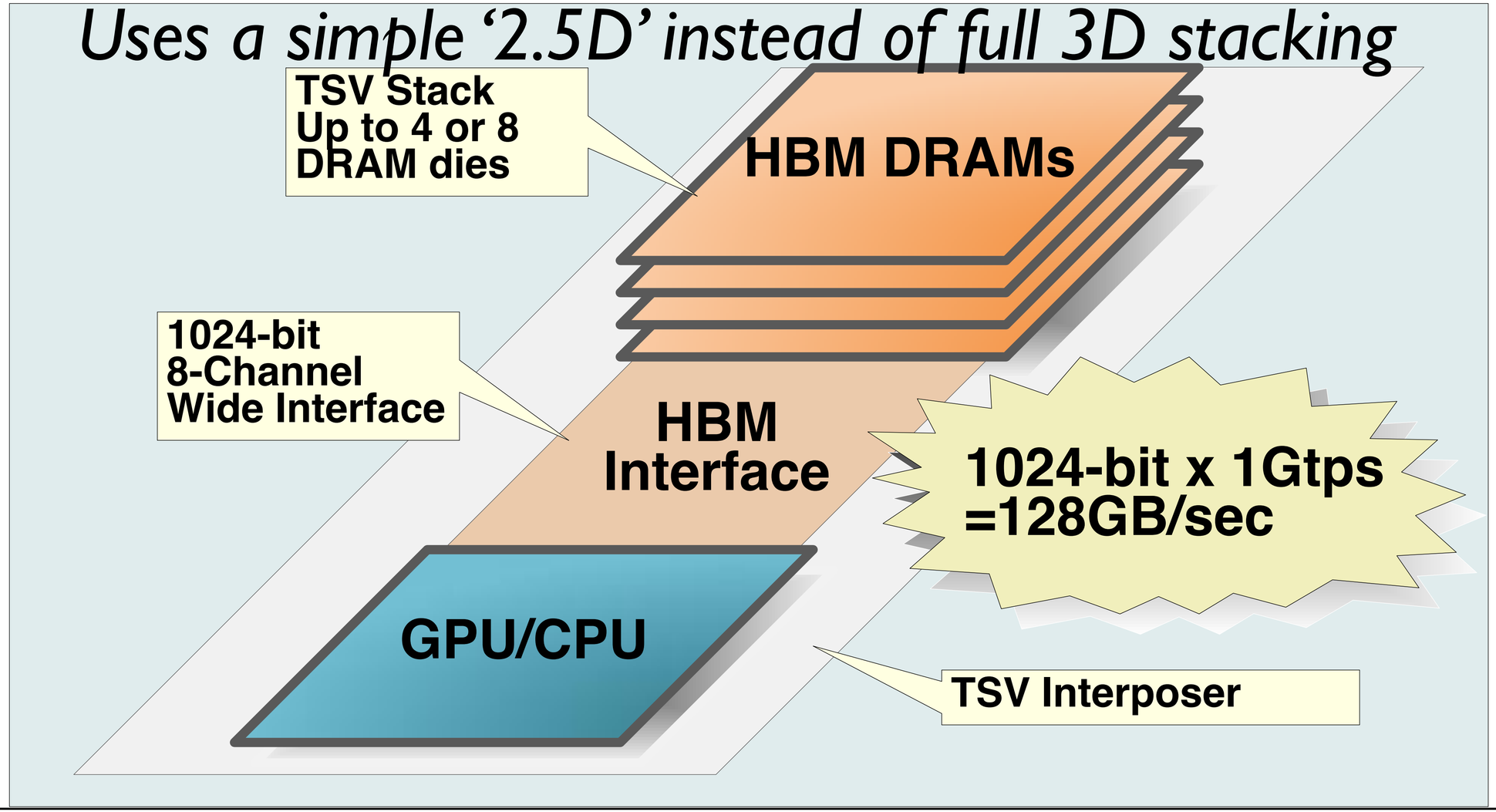
Three Quarters of Ops Memory-Related

Likely Sources of Memory Ops:

Big RF →
- **Capacity Limitations** (increase capacity)

HBM/HMC →
- **Unavoidable** (make local access painless)
- **MPI** (reduce frequency of IPC, reduce cost of sending data over network)

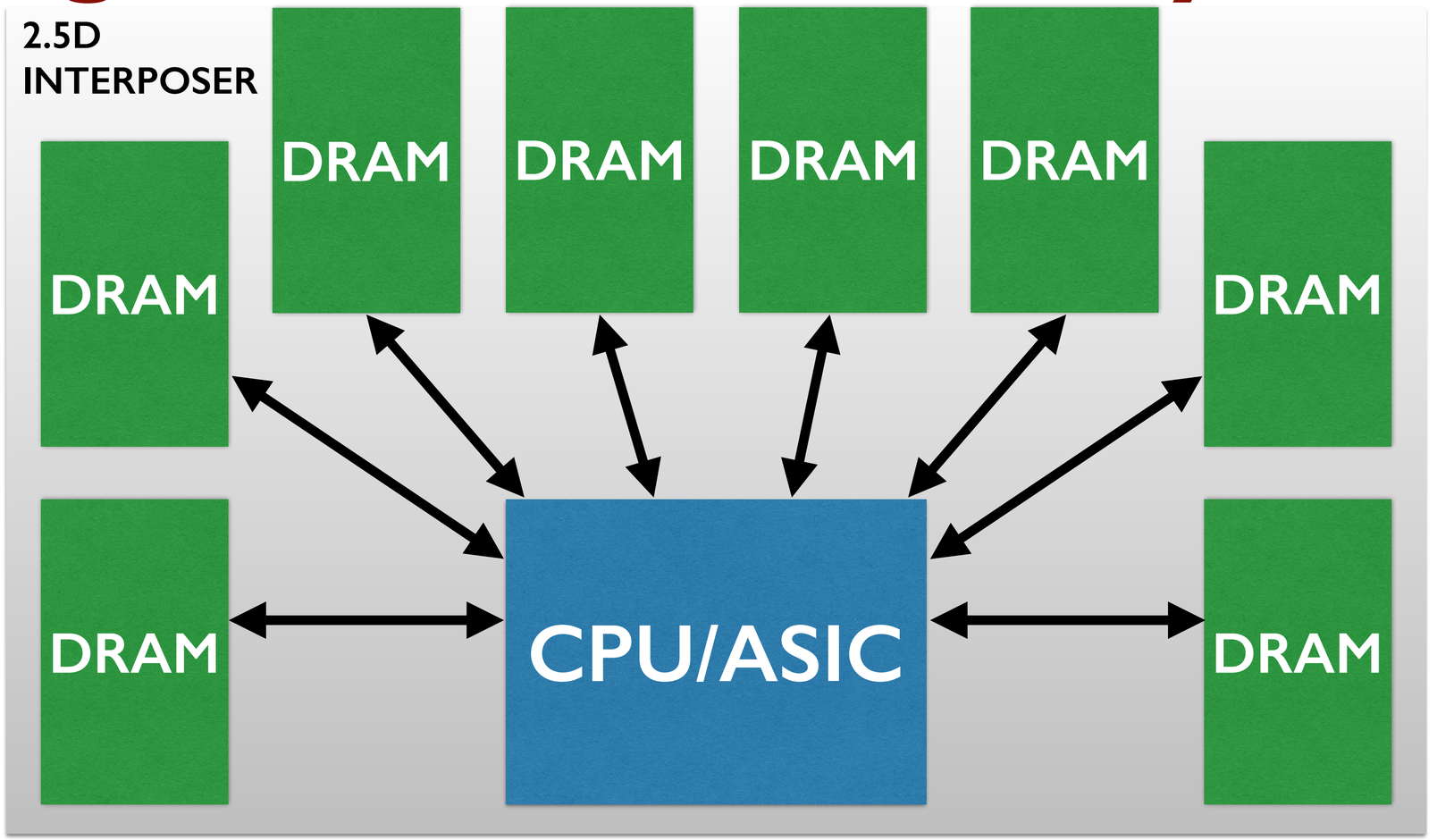
High Bandwidth Memory



Bruce Jacob

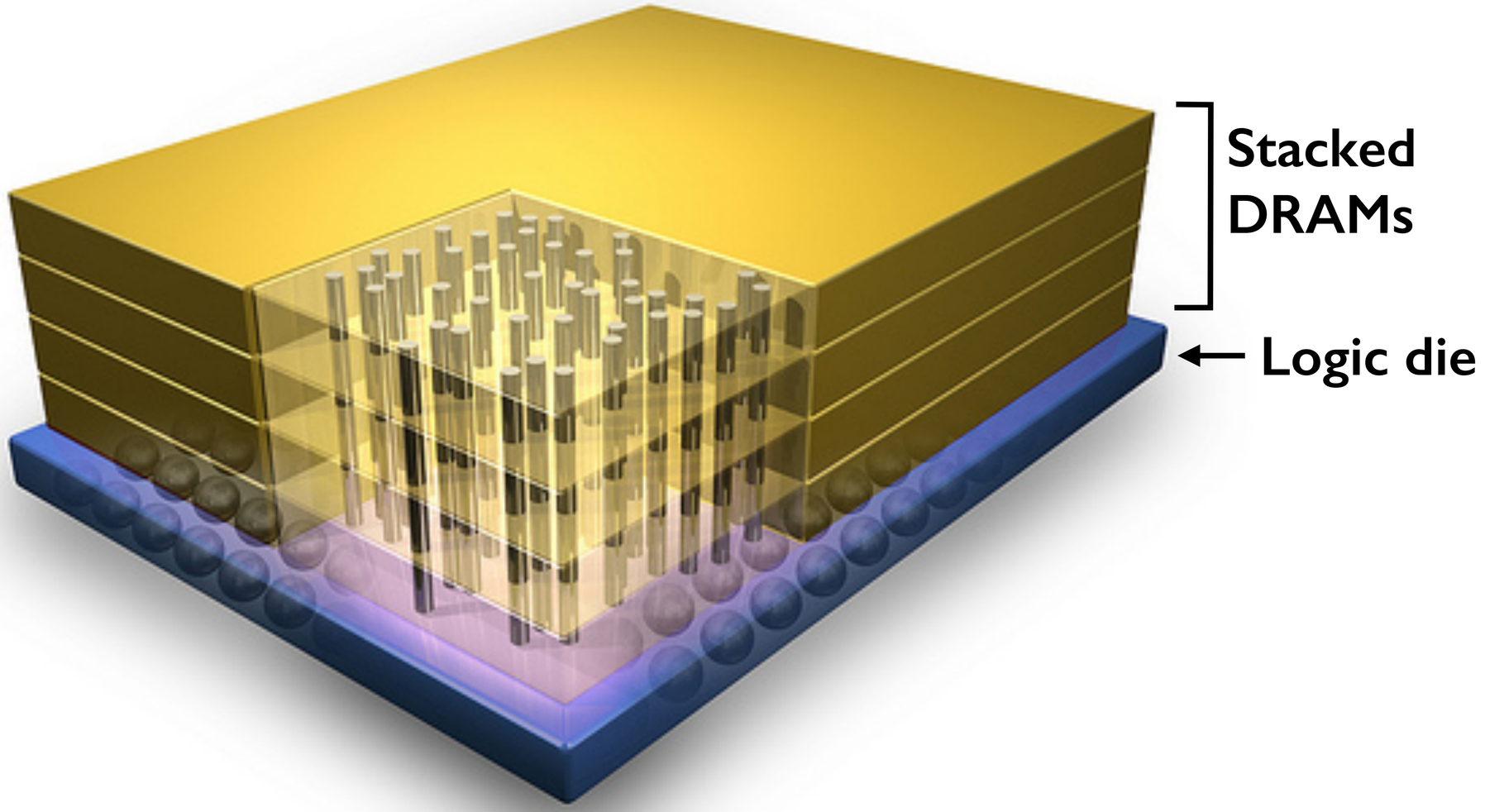
University of
Maryland

High Bandwidth Memory



Each Link is 128 Bits Wide: *1024 Total*

Hybrid Memory Cube



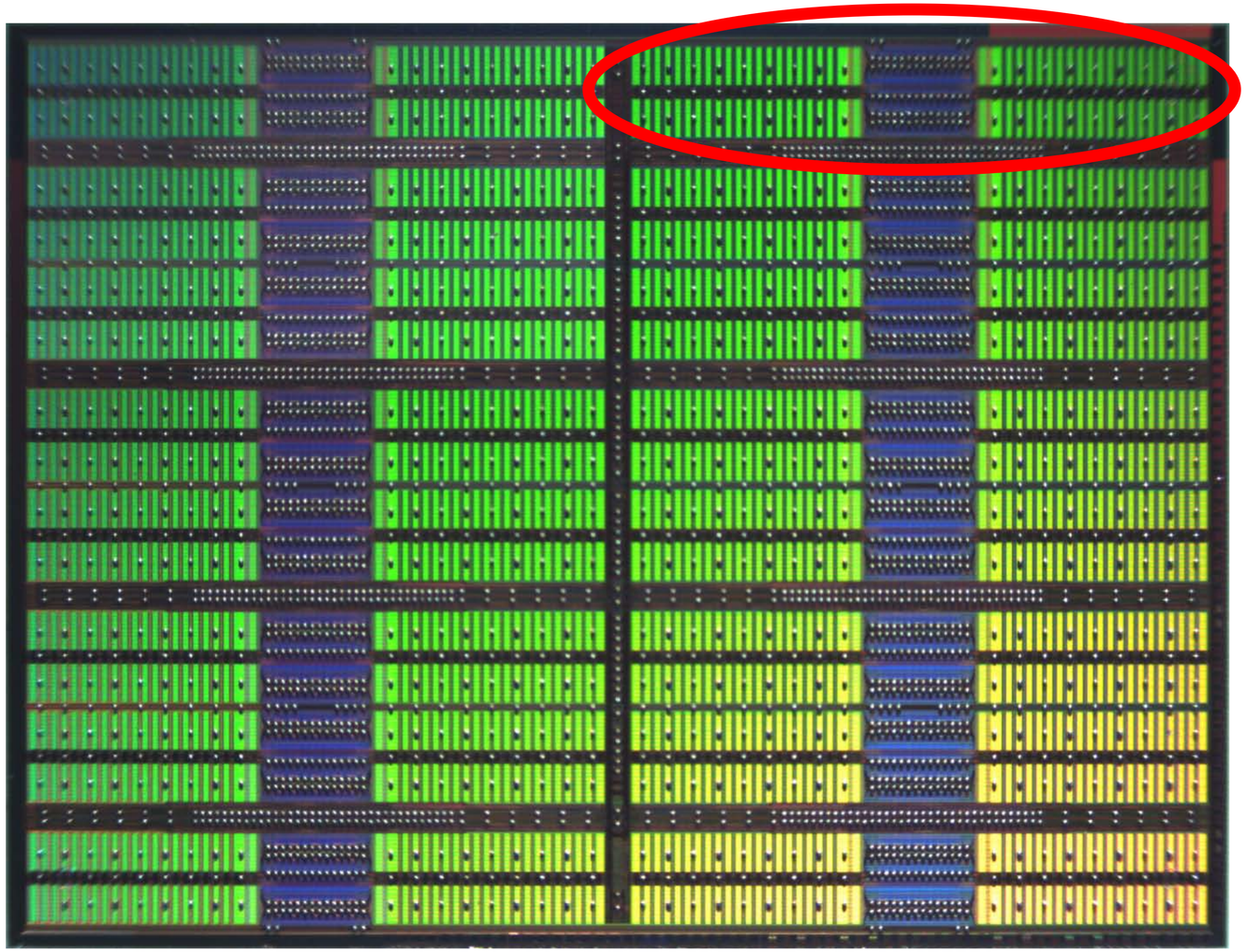
Source: Micron

Bruce Jacob

University of
Maryland

HMC Die
1Gb

Partition, with internal banks

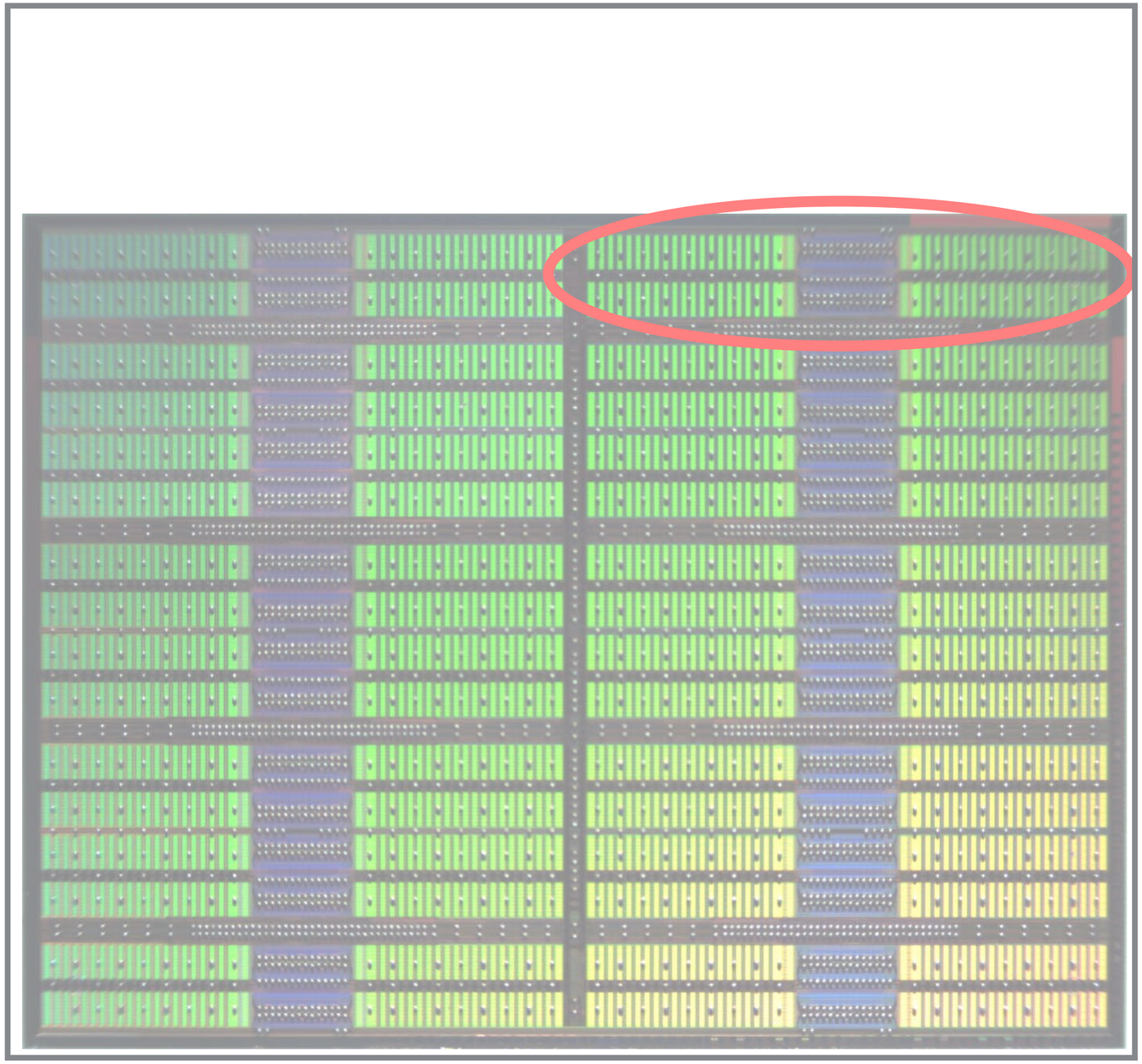


Source: Micron

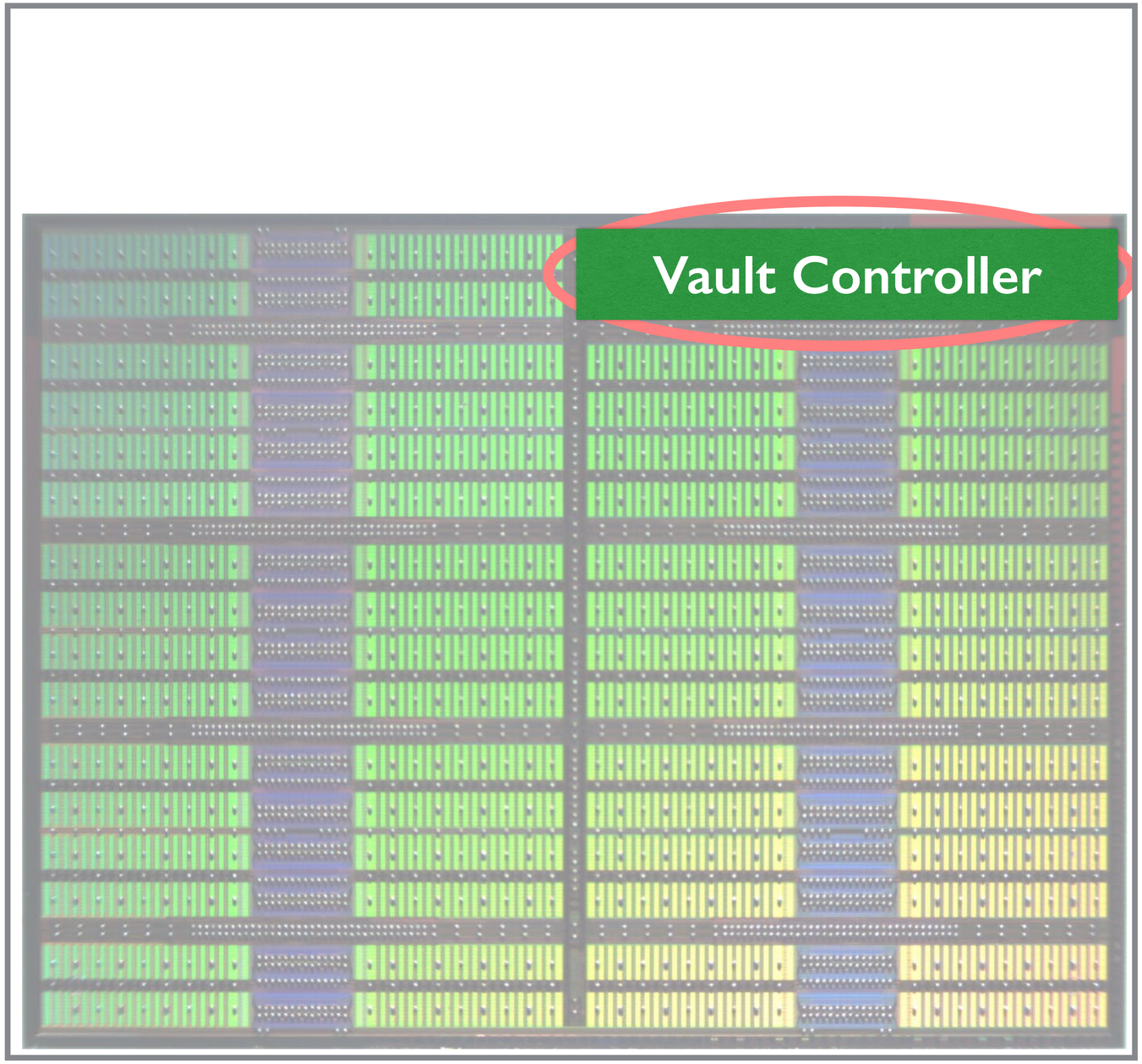
Bruce Jacob

University of
Maryland

Logic Die



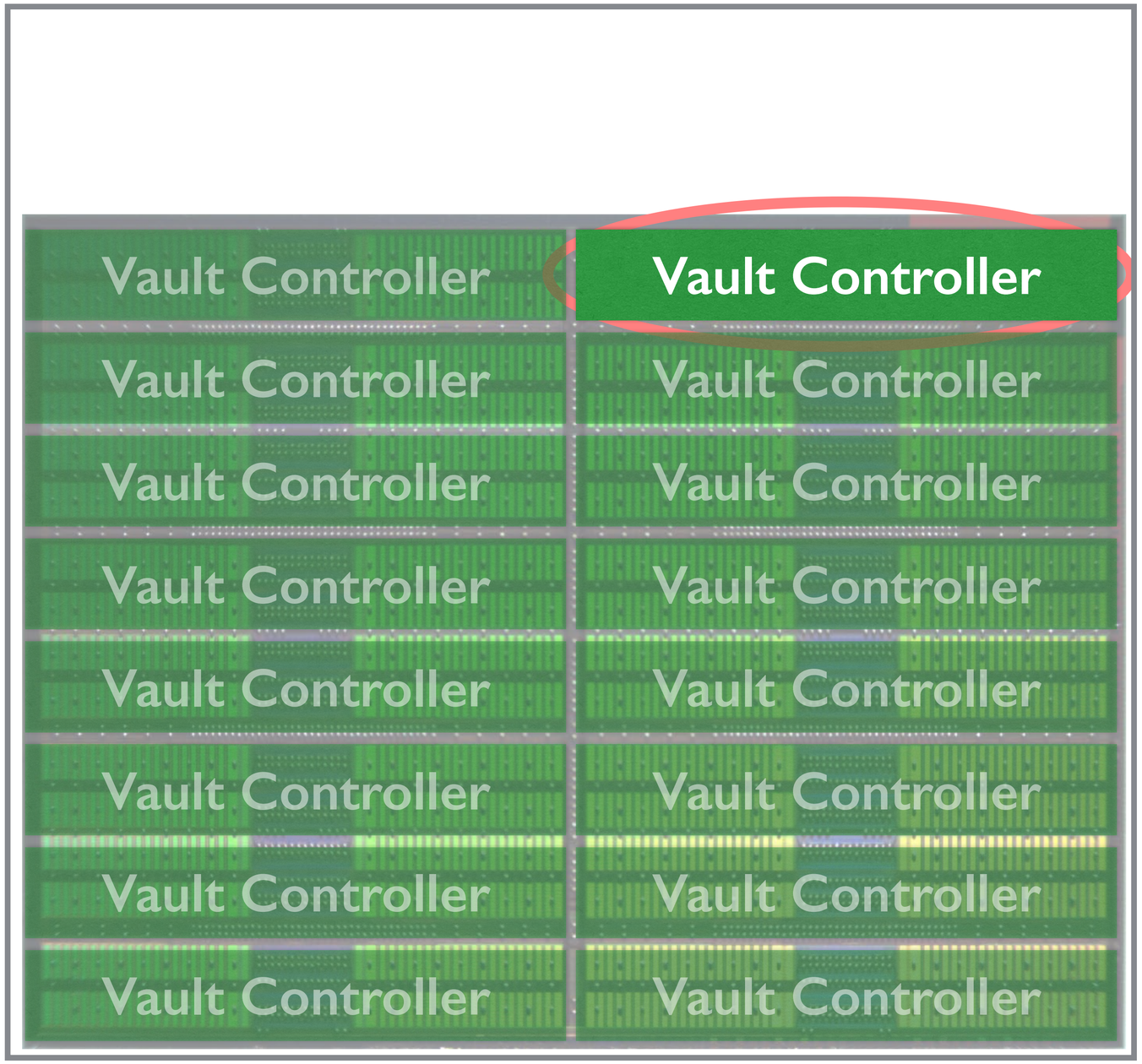
Logic Die



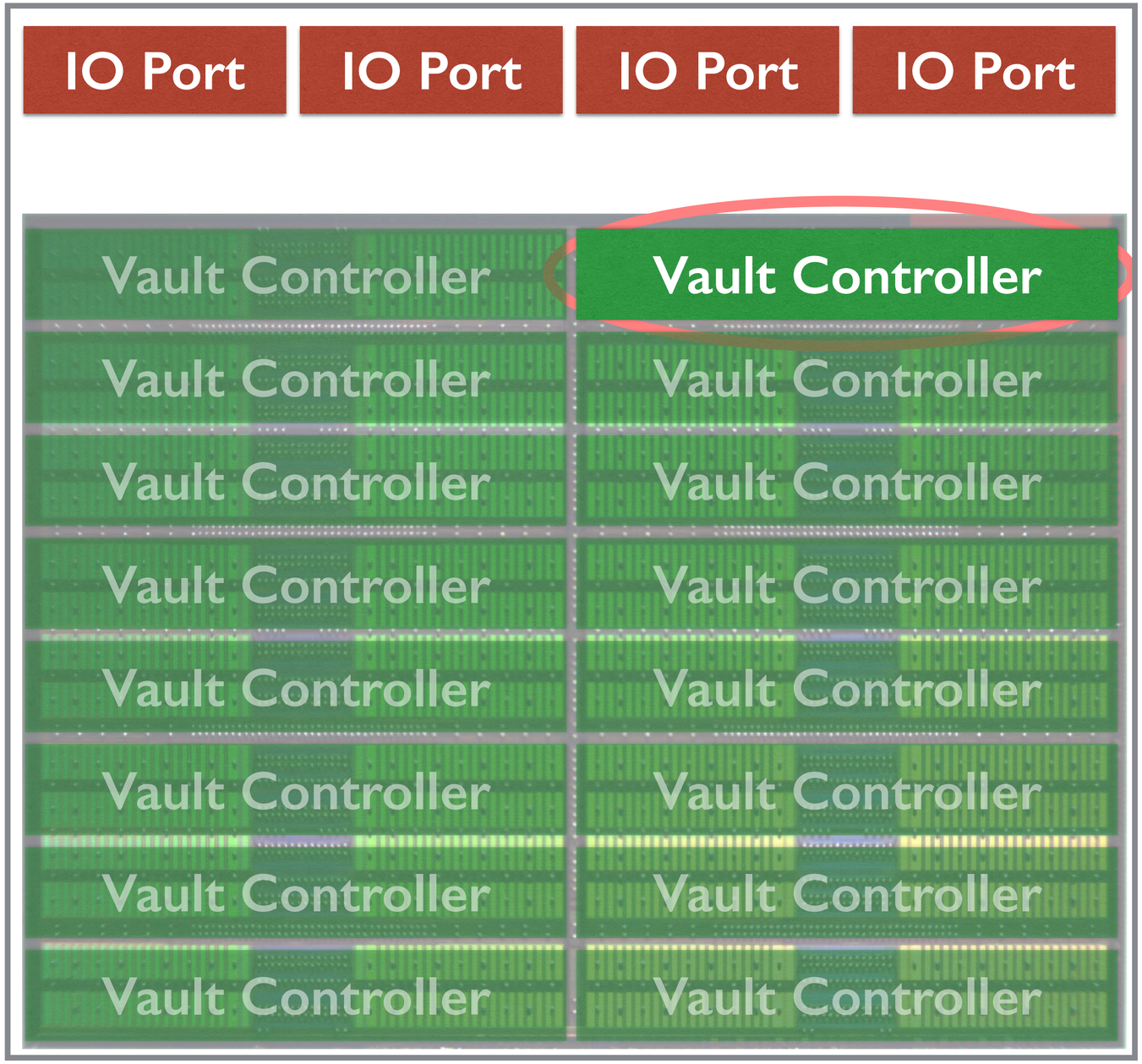
Bruce Jacob

University of
Maryland

Logic Die



Logic Die


IO Port
IO Port
IO Port
IO Port
Vault Controller
Vault Controller
Vault Controller
Vault Controller
Vault Controller
Vault Controller
Vault Controller
Vault Controller
Vault Controller
Vault Controller
Vault Controller
Vault Controller
Vault Controller
Vault Controller
Vault Controller
Vault Controller

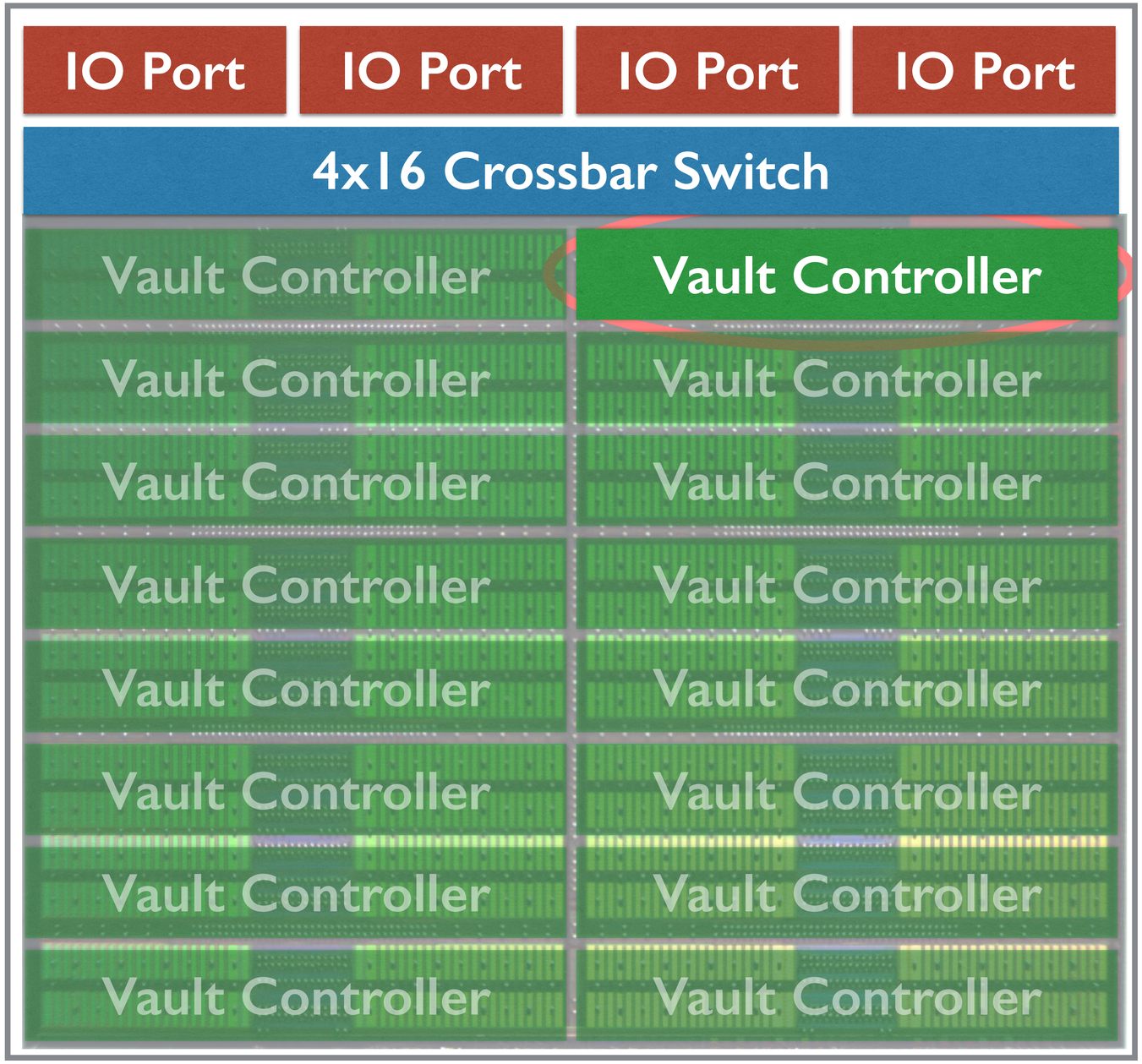
Logic Die

Off-chip: high speed SerDes and generic protocol

I/O Ports up to 80 GB/s each

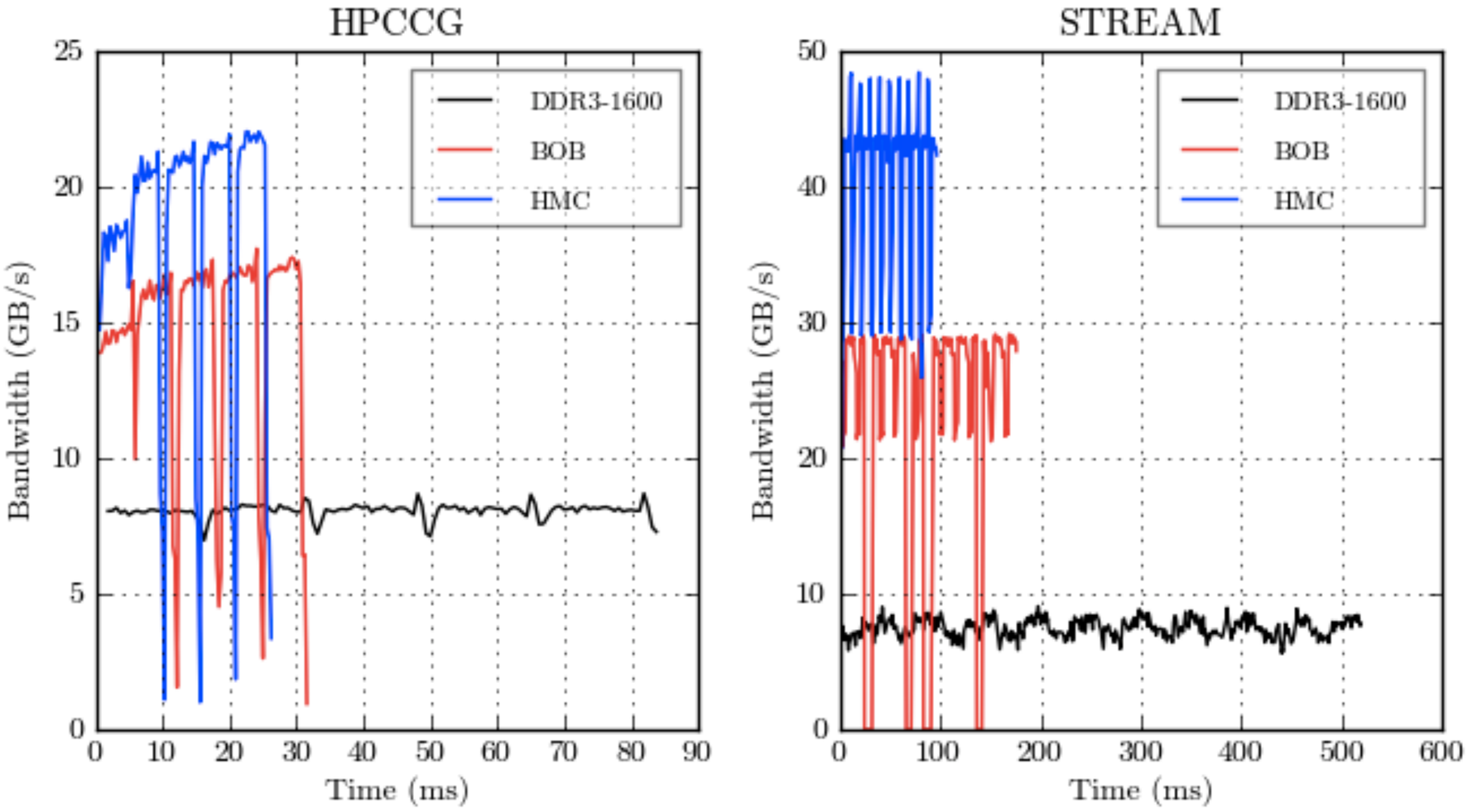
Next gen is twice that

Total conc'y = 16 x 8 x 2..8
(256–1024)



HMC Performance

Execution can be several *times* faster than DDR3-1600



Source: Rosenfeld
Ph.D. Thesis,
U. Maryland 2014

HMC/HBM + Kandou Bus?

Recent ISSCC Talks from Kandou:

- Add CNRZ-5 SerDes to HMC
  => 6 wires, 125 Gbps, 1 pJ/bit @ 28nm
- Add bidirectional "Glasswing" to HBM
  => Reduces pin count from 1700 to 320
  => No interposer necessary
  => 16 Gbaud, 1 pJ/bit @ 28nm

(1000 Gbit/sec @ 1 pJ/bit = 1 J/sec = 1W)

Exascale v. Memory III

Three Quarters of Ops Memory-Related

Likely Sources of Memory Ops:

Big RF →
- **Capacity Limitations** (increase capacity)

HBM/HMC →
- **Unavoidable** (make local access painless)
- **MPI** (reduce frequency of IPC, reduce cost of sending data over network)

Exascale v. Memory III

Three Quarters of Ops Memory-Related

Likely Sources of Memory Ops:

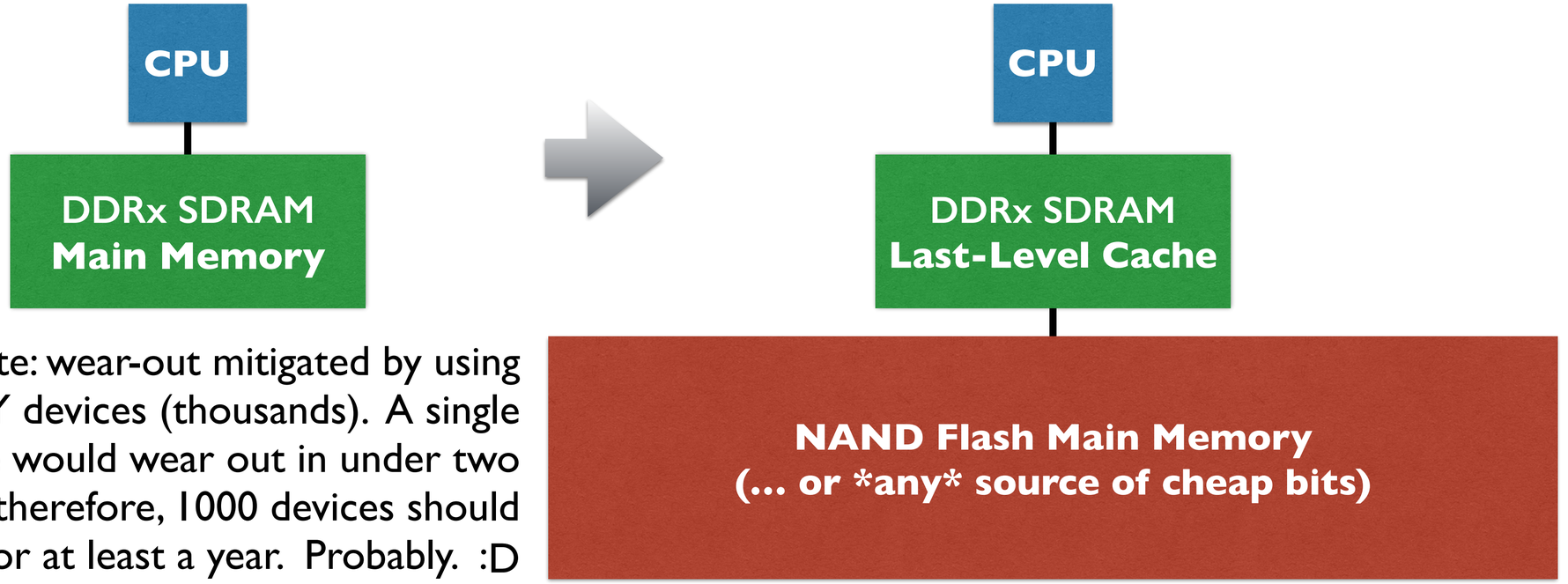
Big RF → • **Capacity Limitations** (increase capacity)

HBM/HMC → • **Unavoidable** (make local access painless)

NVMM → • **MPI** (reduce frequency of IPC,
reduce cost of sending data over network)

Reduce Need to Move Data

	Cost for 10 GB	Size of 10 GB	Power for 10 GB	Power per GB/s
Off-Chip SRAM	$1,000	1 bucket	0.1–1 W	0.1 W
DDR4 SDRAM	$100	1 DIMM	1 W	0.1 W
NAND Flash	$10	1 chip	0	0.1 W (?)
3D XPoint	?	<1 chip	0	0.1 W (?)

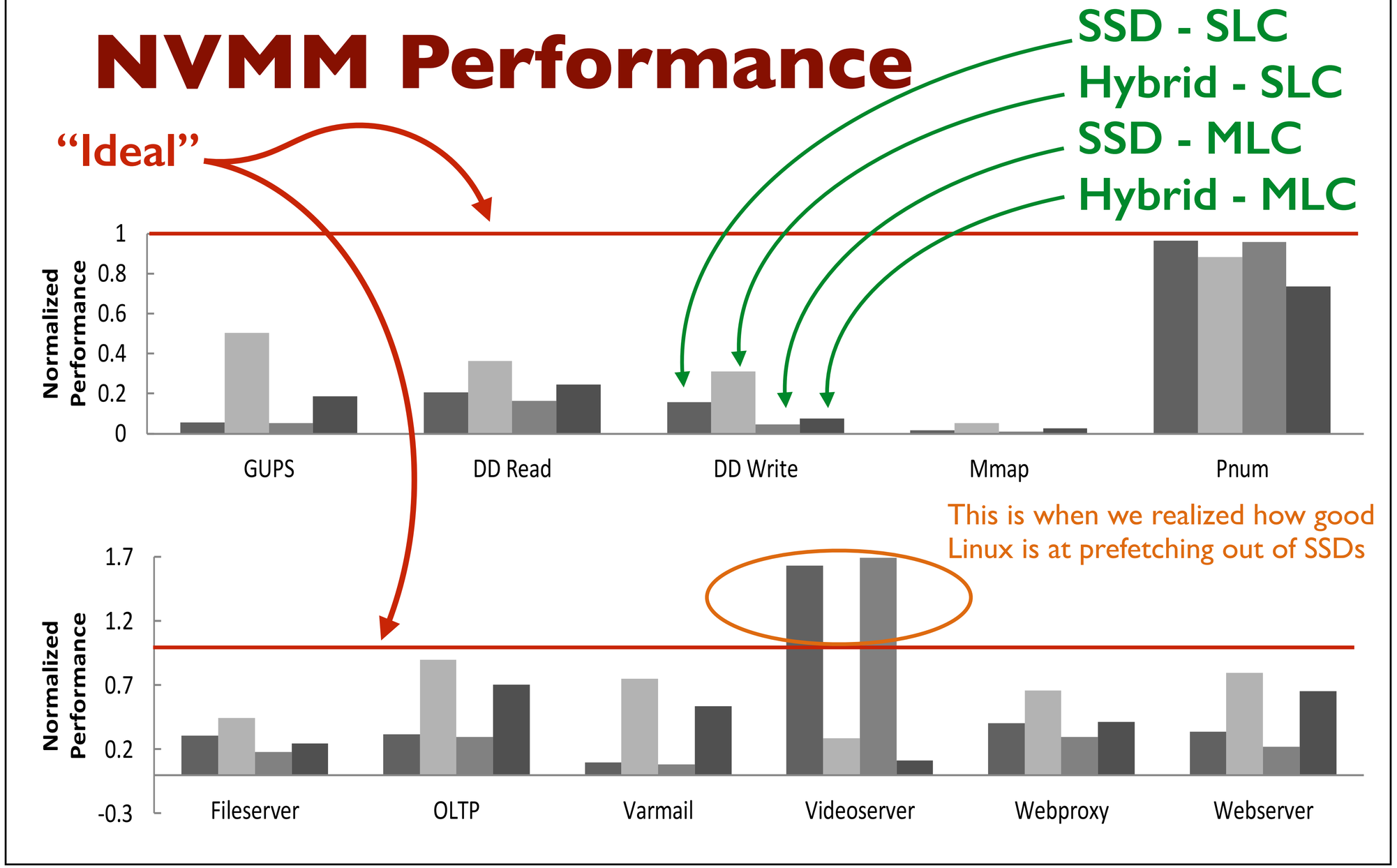


Note: wear-out mitigated by using MANY devices (thousands). A single device would wear out in under two days; therefore, 1000 devices should last for at least a year. Probably. :D

Bruce Jacob

University of
Maryland

NVMM Performance



Exascale v. Memory III

Three Quarters of Ops Memory-Related

Likely Sources of Memory Ops:

Big RF → • **Capacity Limitations** (increase capacity)

HBM/HMC → • **Unavoidable** (make local access painless)

NVMM → • **MPI** (reduce frequency of IPC,
reduce cost of sending data over network)

Exascale v. Memory III

Three Quarters of Ops Memory-Related

Likely Sources of Memory Ops:

Big RF → • **Capacity Limitations** (increase capacity)

HBM/HMC → • **Unavoidable** (make local access painless)

NVMM → • **MPI** (reduce frequency of IPC,

Fishnet → reduce cost of sending data over network)

Fishnet Interconnect

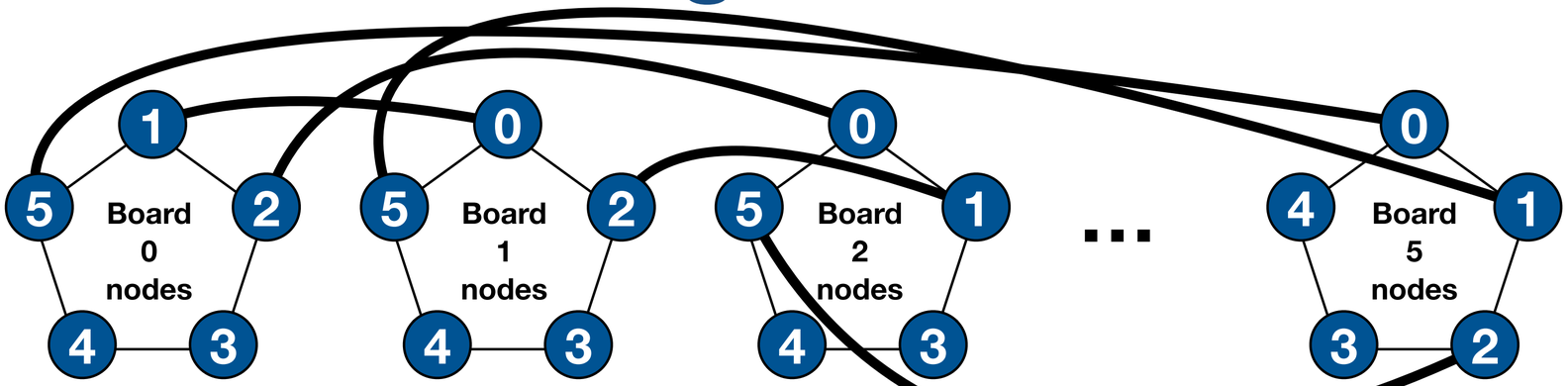
Lite:

For all i,j:
subnet i, node j
connects to
subnet j, node i

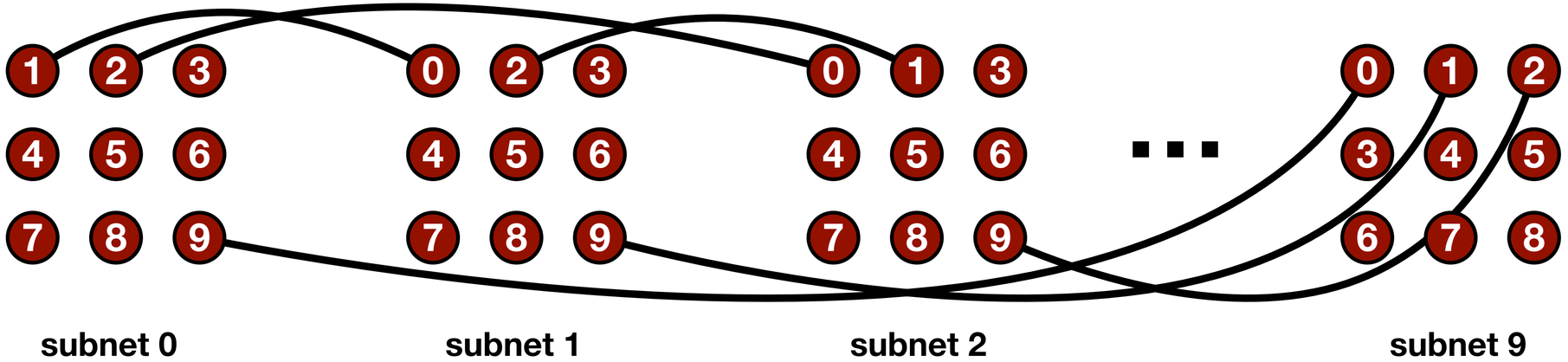
Regular:

For all i,j:
subnet i, node j
neighbors
connect to
subnet j, node i
neighbors

Fishnet + Moore → **Angelfish:**



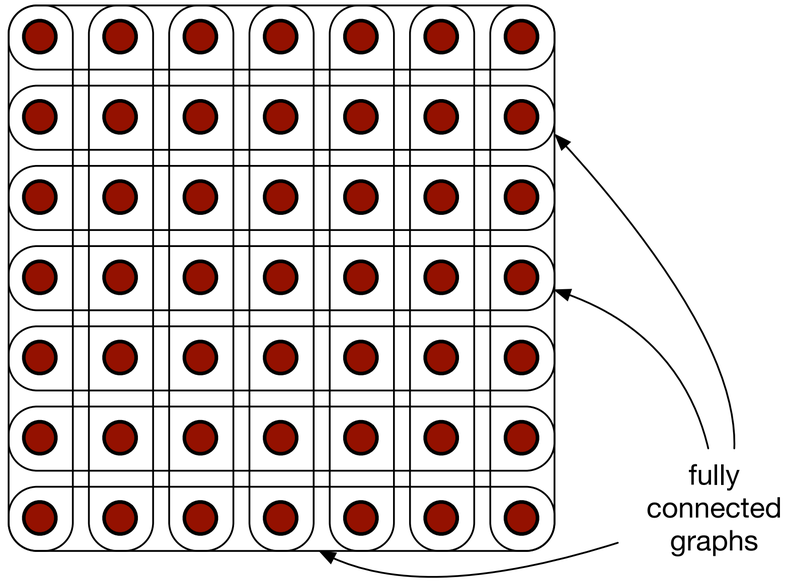
+ Flat.Butterfly → **Flying Fish:**

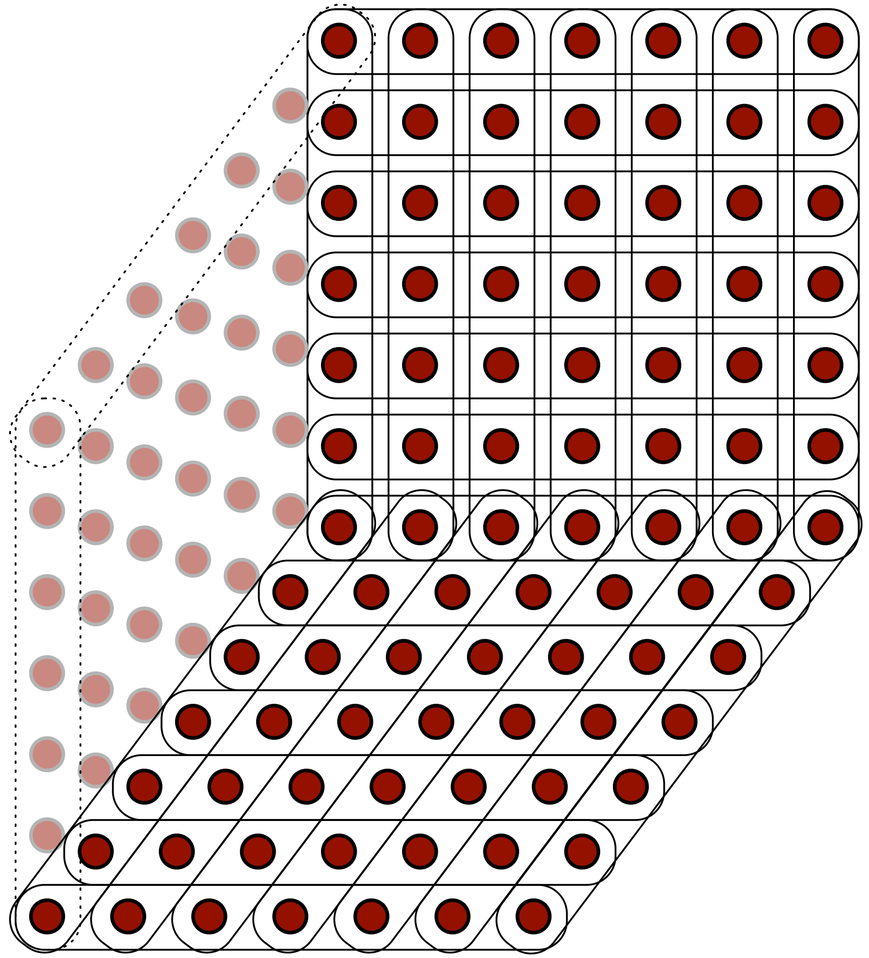


Bruce Jacob

University of
Maryland

High-D FButterfly and Torus





Nodes: n^D
Ports: D(n–1)
Latency: D

Bruce Jacob

University of
Maryland

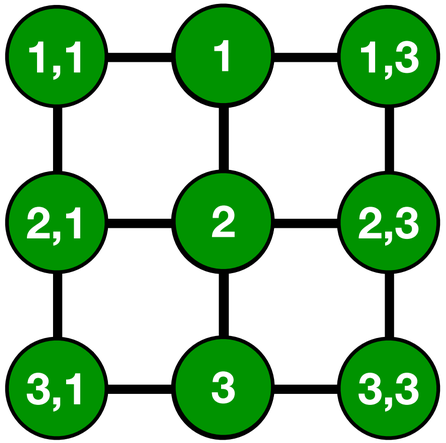
High-D FButterfly and Torus

Nodes: n^D, Ports: 2D, Latency: variable

Bruce Jacob

University of
Maryland

High-D FButterfly and Torus

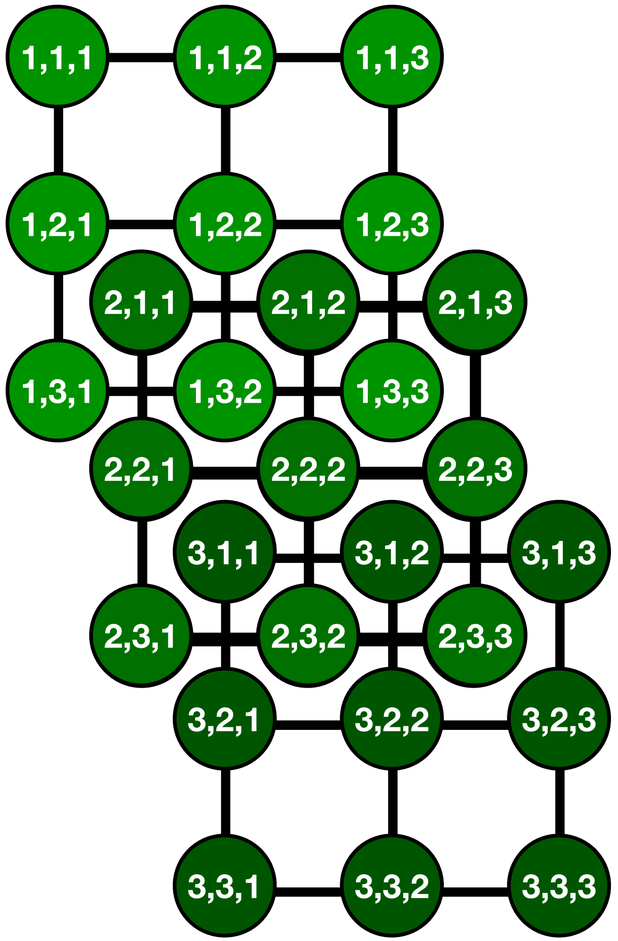


Nodes: n^D, Ports: 2D, Latency: variable

Bruce Jacob

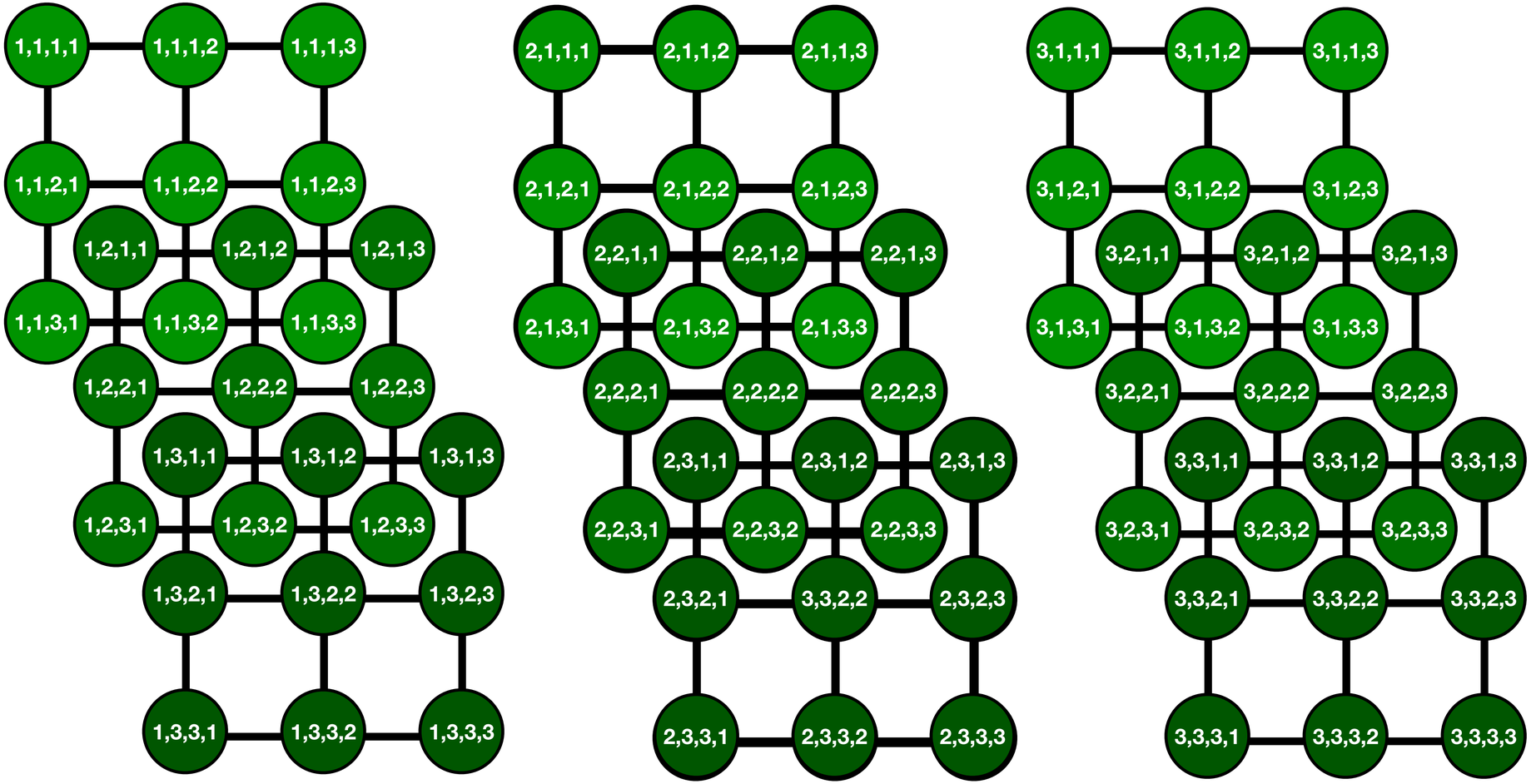
University of
Maryland

High-D FButterfly and Torus



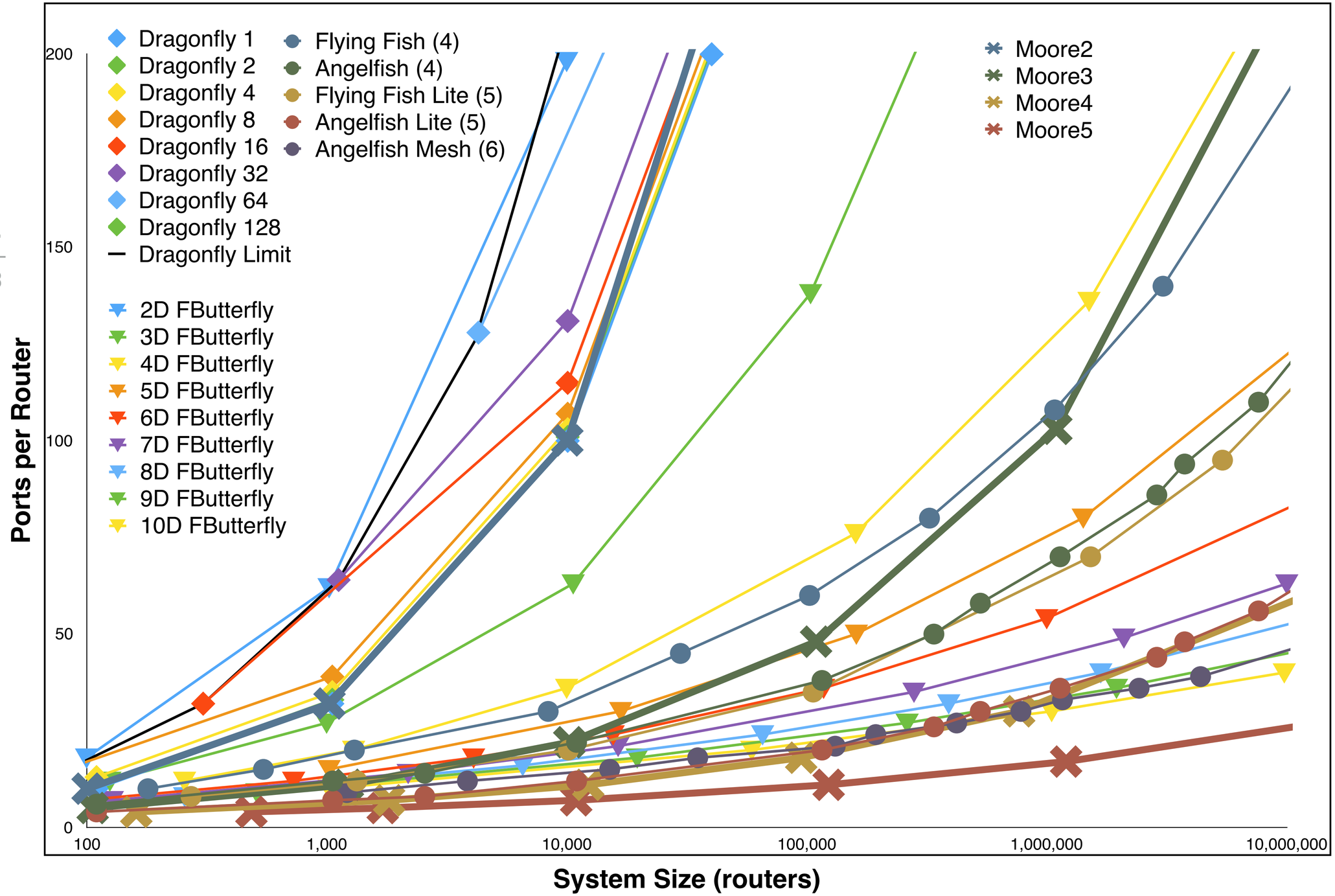
Nodes: n^D, Ports: 2D, Latency: variable

High-D FButterfly and Torus



Nodes: n^D, Ports: 2D, Latency: variable

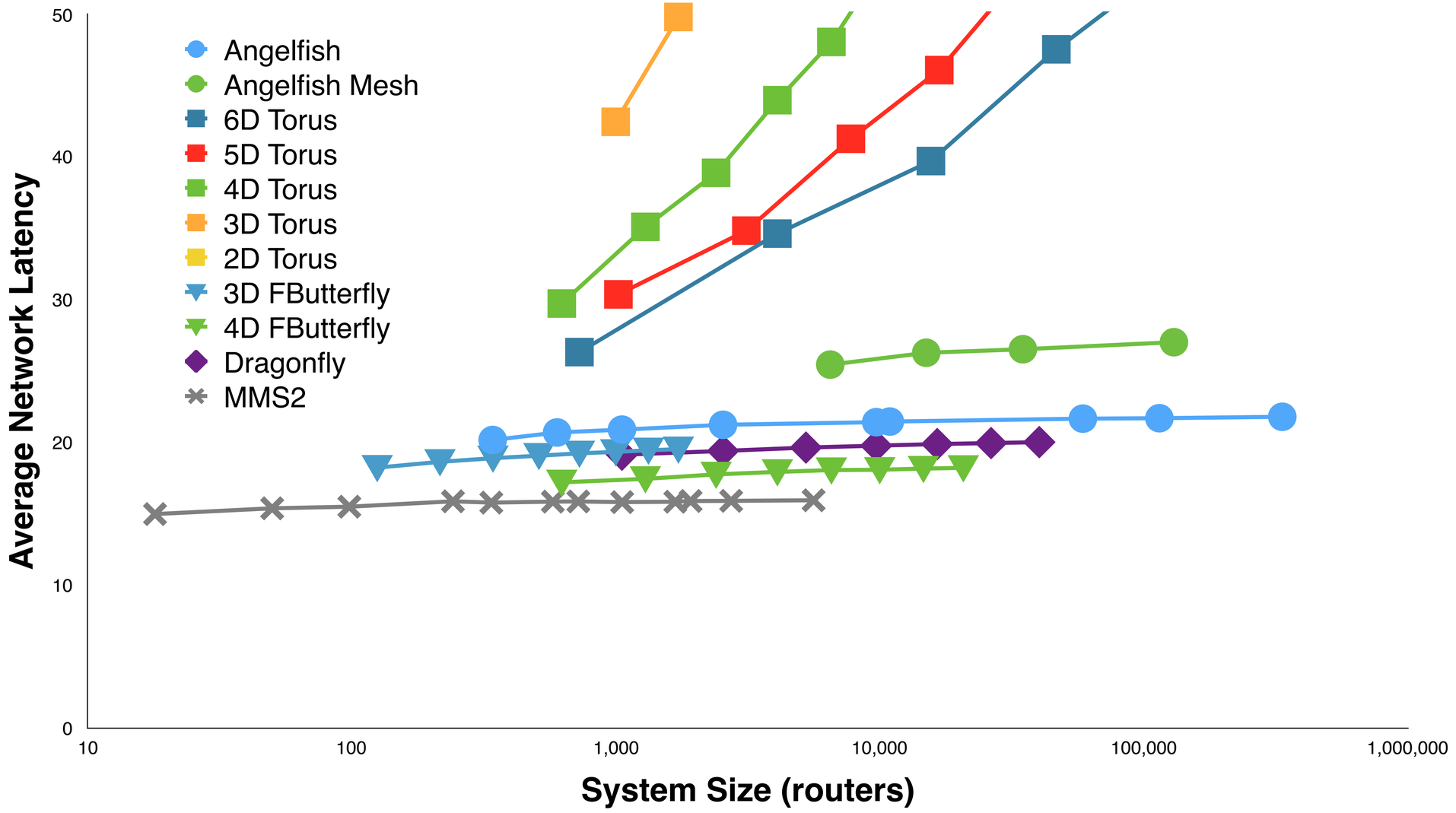
Port Costs


Dragonfly 1
Dragonfly 2
Dragonfly 4
Dragonfly 8
Dragonfly 16
Dragonfly 32
Dragonfly 64
Dragonfly 128
Dragonfly Limit
Flying Fish (4)
Angelfish (4)
Flying Fish Lite (5)
Angelfish Lite (5)
Angelfish Mesh (6)
Moore2
Moore3
Moore4
Moore5
2D FButterfly
3D FButterfly
4D FButterfly
5D FButterfly
6D FButterfly
7D FButterfly
8D FButterfly
9D FButterfly
10D FButterfly
Ports per Router
0
50
100
150
200
100
1,000
10,000
100,000
1,000,000
10,000,000
System Size (routers)

Bruce Jacob

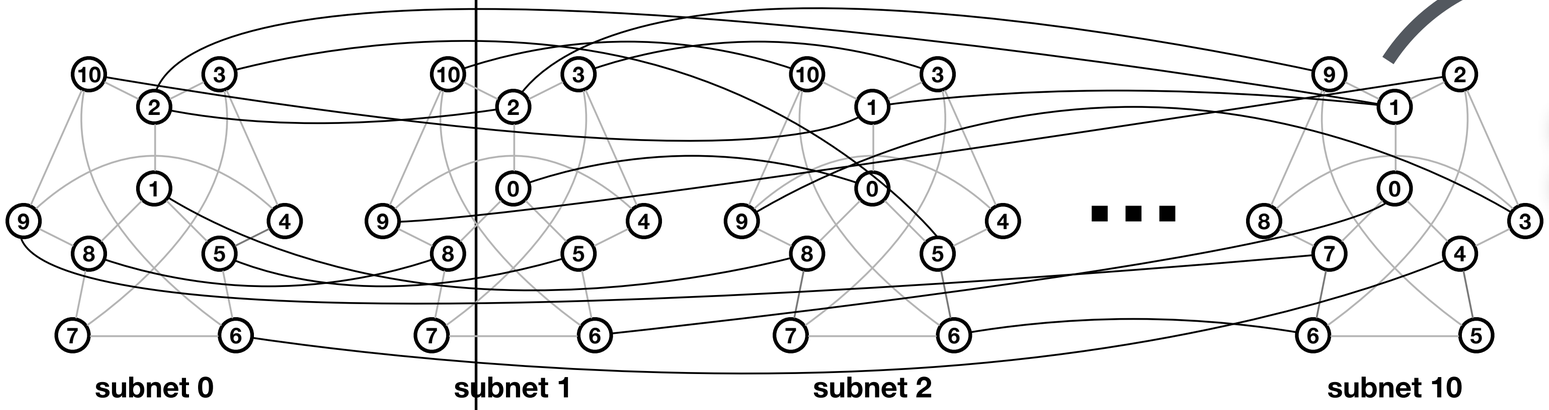
University of
Maryland

Cycle-Accurate Simulations

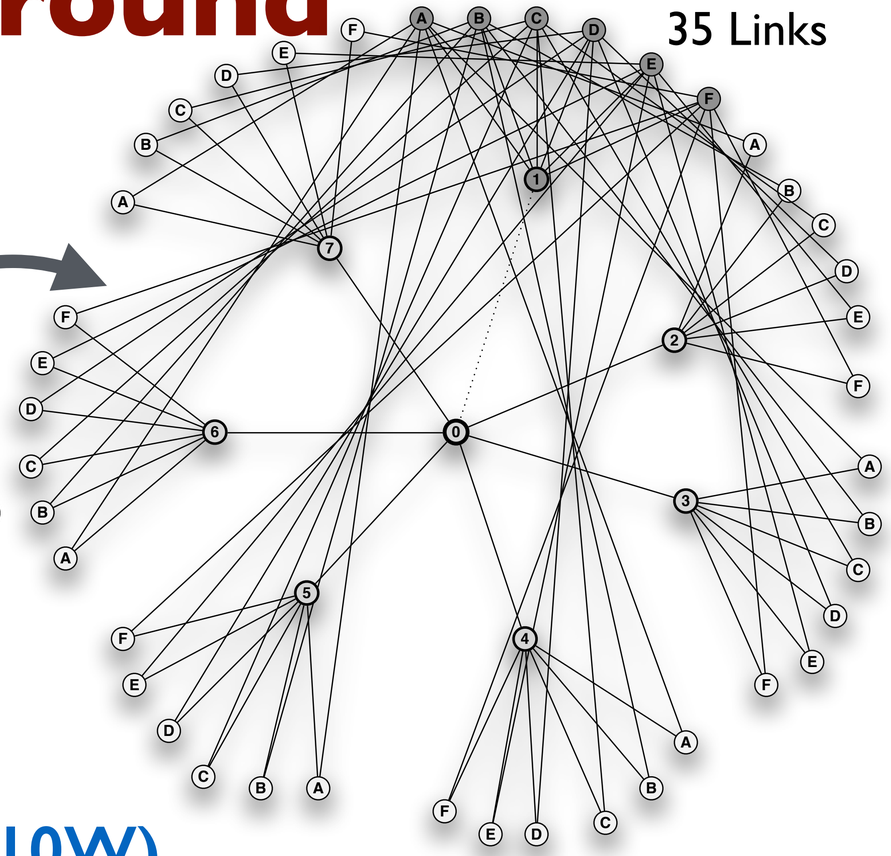


A Stake in the Ground

Angelfish 1059x1058



(Real) Moore Graph
1058 Nodes, Diameter 2
35 Links



1,120,422 nodes:

- Low-Power CMP (~10W)
- 1 TB NV Memory (~3W)
- 32 GB Hi-BW DRAM (~6W)
- 70 I/O links @ 1GB/s (~7W)

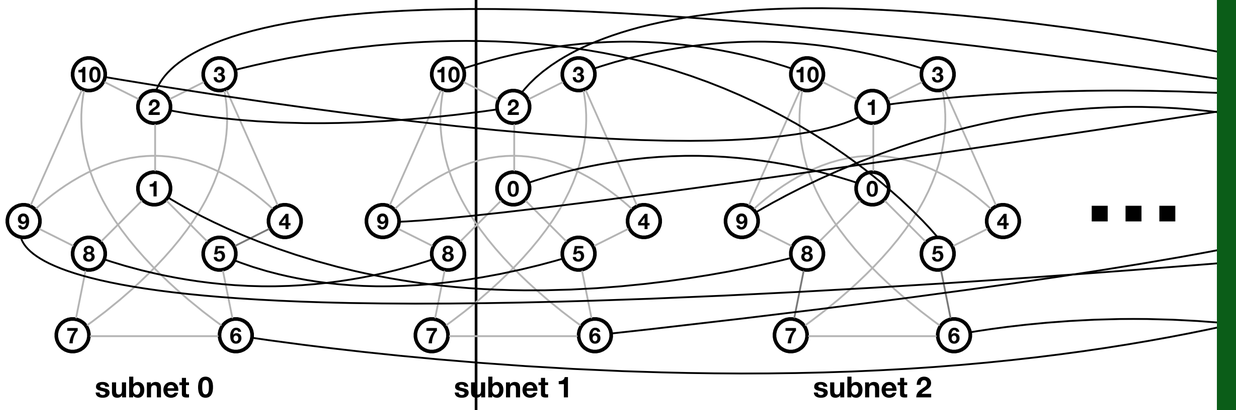
Bruce Jacob

University of
Maryland

A Stake in the Ground

(Real) Moore Graph
1058 Nodes, Diameter 2
35 Links

Angelfish 1059



...

System-Level:
1.1 M nodes, Diameter 4
1.1 Exabyte Main Memory
10 PB/s Bisection Bandwidth
1/2 – 3/4 ExaFLOP (peak)
29 MW

1,120,422 node

- Low-Power CMP (~10W)
- 1 TB NV Memory (~3W)
- 32 GB Hi-BW DRAM (~6W)
- 70 I/O links @ 1GB/s (~7W)

Recap, Bottom Line

Reduce Main-Memory Accesses: **Big RF**

Reduce the Cost Per: **High-BW DRAM**

Reduce Remote Accesses: **Big, LP MM**

Reduce the Cost Per: **Lo-D,Hi-BW Net**

Memory is the problem …
fix it, you get exascale <u>way</u> before 2023

Acknowledgments

Sandians: Jim Ang, Janine Cook, Scott Hemmert, Dave Resnick, Arun Rodrigues

Terps (past & present): David Banks, Ishwar Bhati, Mu-Tien Chang, Elliott Cooper-Balis, Max DePalma, Ahmed Elshaarany, Po-Chun Huang, Shang Li, Paul Rosenfeld, Emily Ruppel, Jim Stevens, Paul Tschirhart, Yitian Wang

Support: DOD, DOE, Intel, NGC, Sandia

Shameless Plug

www.memsys.io

Thank You!

Bruce Jacob
University of Maryland

blj@umd.edu
www.ece.umd.edu/~blj

MEMSYS 2016

The International Symposium on Memory Systems — October 3–6, Washington, DC

Important Dates

submission, up to 15 pages

Organizers

Bruce Jacob, U. Maryland
Kathy Smiley, Memory Systems

Ameen Akel, Micron
James Ang, Sandia National Labs
Yitzhak Birk, Technion
Bruce Childers, U. Pittsburgh
Zeshan Chishti, Intel
Chen Ding, U. Rochester
David Donofrio, Berkeley Lab
Wendy Elsasser, ARM
Maya Gokhale, LLNL
Thuc Hoang, NNSA/DOE
Hillery Hunter, IBM
Mike Ignatowski, AMD
Aamer Jaleel, NVIDIA
David Kaeli, Northeastern
Scott Lloyd, LLNL
Gabriel Loh, AMD
Kenneth Ma, Hynix
Richard Murphy, Micron
Mike O'Connor, NVIDIA
Petar Radojkovic, BSC
David Resnick, Sandia National Labs
Arun Rodrigues, Sandia National Labs
John Shalf, Berkeley Lab
Anouk Van Laer, U. College London
Jeffrey Vetter, Georgia Tech & ORNL
Robert Voigt, Northrop Grumman
David Wang, Inphi
Christian Weis, U. Kaiserslautern
Kenneth Wright, Rambus
Sudhakar Yalamanchili, Georgia Tech

The memory system has become extremely important recently: memory is slow, and this is the primary reason that computers don't run significantly faster than they do. In large-scale computer installations ·sized systems powering Google.com, Amazon.com, tor, memory is often the largest dollar cost as well ner of energy. Consequently, improvements in the have significant impact on the real world, d energy, performance, and/or dollar cost. of the problems we see in the memory system are nature—their solution would likely require work at cations to circuits. Thus, while the scope of the the scope of the solutions will be much wider.

t

ıed papers containing significant novel ideas and technical results are solicited. Papers that focus on **system, software, and architecture level concepts**, outside of traditional conference scopes, will be preferred over others (e.g., the desired focus is away from pipeline design, processor cache design, prefetching, data prediction, etc.). Symposium topics include, but are not limited to, the following:

- Memory system design from both hardware & software perspectives
- Operating system design for hybrid/nonvolatile memories
- Technologies including PCM, flash, DRAM, STT-RAM, 3DXP, etc.
- Data-movement issues and mitigation techniques
- Interconnects to support large-scale data movement
- Software & application techniques for distributed memories
- Algorithmic & software memory-management techniques
- Near-memory computing
- Memory-centric programming models & compiler techniques
- Memory failure modes and mitigation strategies
- Memory and system security issues

To reiterate, papers that focus on topics outside of traditional conference scopes will be preferred over others.

Submissions and Presentations

Our primary goal is to showcase interesting ideas that will spark conversation between disparate groups—to get applications people and operating systems people and system architecture people and interconnect people and circuits people to talk to each other. We accept extended abstracts, position papers, and/or full research papers, and each accepted submission is given a 20-minute presentation time slot. **All accepted papers will be published in the ACM Digital Library.**